IDIAP Research Report

Towards using slide information to enhance speech transcription of meetings

Artem Peregoudov [a] [b]

Alessandro Vinciarelli [a] [b] Hervé Bourlard [a] [b]

IDIAP–RR 06-01

January 2006

submitted for publication

[a] IDIAP Research Institute, Martigny, Switzerland

[b] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**

Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland

Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-01

Towards using slide information to enhance speech transcription of meetings

Artem Peregoudov Alessandro Vinciarelli Hervé Bourlard

January 2006

submitted for publication

Abstract. In this paper we investigate the possibility of improving the speech recognition performance of meeting recordings by using slides captured during the recording process. The key hypothesis exploited in this work is that both slides and speech carry correlated contextual and semantic information. Thus, we propose an approach using the information extracted from slides aimed at reducing the speech recognition word error rate. The N-Best lists output by the recogniser are rescored through Information Retrieval techniques to maximise the similarity between speech and slides transcripts. Results obtained on three meeting recordings (for a total duration of about 90 minutes) show no statistically significant variation of the word error rate. Additional studies provide further insight based on both language properties and statistics of the word distributions in the two sources.

1 Introduction

Recognition of spontaneous speech is still a major challenge for current Automatic Speech Recognition (ASR) systems. While the Word Error Rate (WER) is around 5-10% on constrained (or planned) speech, such as the one of broadcast news or dictation systems [1], it increases up to 30-40% on spontaneous speech, such as the one of meetings, involving speaker hesitations, interruptions and various accents [2][3]. However, state-of-the-art systems are based on the speech signal only while it is frequent to process multimedia recordings where additional information, potentially helpful to the recognition, is available.

In this paper, we focused on meetings recorded in a so-called *Smart Meeting Room* [4], i.e. a room equipped with capturing devices such as videocameras, microphone arrays, framegrabbers, tablets, etc. All devices are synchronised and allow the recording of audio, vusual and textual streams within meeting scenarios.

The key assumption made in this work is that slides and speech carry the same contextual/semantic information, thus, we focused on two streams: the first one is the speech transcription obtained from a Large Vocabulary Continuous Speech Recognition (LVCSR) system, the second one is the text automatically extracted from the slides (see Section 2.2 for more details).

The speech recognition is carried out by a state-of-the-art LVCSR system [2] which outputs N-Best lists, i.e. a list of the N most likely transcriptions for a given utterance, ranked with respect to their likelihood scores. The sequence of N-Best lists is synchronised with the slides (Figure 2), so that it is possible to rescore the hypotheses from the N-Best lists based on the content of the slide(s) displayed during the corresponding time interval. Such a N-Best list rescoring technique is expected to improve the recognition rate.

We applied the above method to a test set containing three meetings, of a total duration of about 90 minutes corresponding to approximately 15,000 words. The rescoring results show no statistically significant improvements, the WER being around 33% before and after the rescoring process. One possible reason is that, in general, a few words in a text are subject-dependent. This observation applies to both slides and ASR transcriptions, and even if slides are expected to carry subject-specific information, the rescoring will affect only a few terms in the ASR transcription and thus yield a minor variation of the WER. Such an explanation is supported by the Zipf's law (see Section 3.1) and by the statistical independence between slides displayed and words uttered (see Section 3.2 for more details).

2 Experimental Setup

This section describes how the two streams of interest, the speech and the slides, are captured, presents in more detail the rescoring method and provides a description of the data used in the experiments.

2.1 Acquisition System

The data is captured online during the meetings, as illustrated in the block diagram on Figure 1. All meeting participants are equipped with headsets and lapel microphones which are used to acquire the speech signal. The slides projected on the screen through the PC-projector are captured with a framegrabber, i.e. a hardware device which stores as still images what is displayed on the screen at regular time intervals (e.g. once per second). The resulting image sequence is segmented automatically in correspondence of slide changes.

A state-of-the-art LVCSR system (with a vocabulary of 50k words) [2] is used to transcribe the speech in the form of N-Best lists, while an advanced OCR system [5], robust to complex background and font variability, is used to detect and recognise the text in the slide images. The advantages of such an approach with respect to the use of electronic versions of the slides are shown in [6].

Both streams (utterances/N-Best lists and slide transcriptions) are timestamped during the recognition process resulting in two sequences, subject to recognition errors: a sequence of I utterances $u^{(i)}$ synchronised with a sequence of J slides $s^{(j)}$. Thus, the slide $s^{(j)}$ displayed during every utterance

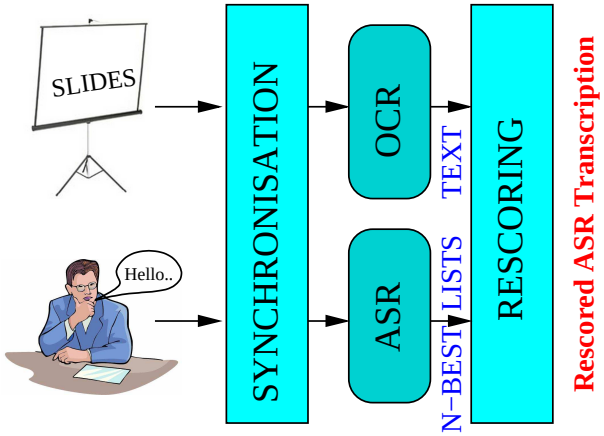

Figure 1: Block diagram of the workflow.

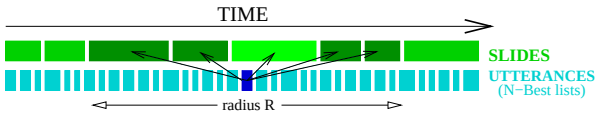

Figure 2: Synchronised sequences of slides and utterances (N-Best lists). The rescoring of an utterance takes into account a slide context defined by a radius R ($R = 2$ in this example).

$u^{(i)}$ is known and can be used to perform the rescoring (see Section 2.2). The slide $s^{(j)}$ is the center of a sequence $C = \{s^{(j-R)}, \ldots, s^{(j)}, \ldots, s^{(j+R)}\}$ that can be called *slide context* of $u^{(i)}$ (see Figure 2).

2.2 Rescoring

The text extracted from the slides is used to rescore the N-Best lists output by the recogniser. Assuming some correlation exists between slide content and speech transcription, it is possible to find the transcription of the N-Best list that better fits the content of the slide displayed at the same moment through similarity measures applied in Information Retrieval (IR) [7] or topic detection [8]. Similarly to previous works in those fields [7][8][9], the utterances and the content of the slides will be represented as vectors and the cosine of the angle between such vectors will be used as a similarity function.

The first step of the process is the so-called *stopping* i.e. removing all words (the so-called *stop-words*) supposed to be content neutral like articles, prepositions, verbs of common use (e.g. *to be* and *to have*), etc. In our experiments, we used a *stoplist* (i.e. a set of stopwords) of 384 words and, after the stopping, the number of words in both speech and slide transcriptions is reduced by about 55% (see Table 3). After stopping, the content of the N-Best lists and of the slides is used to extract the so-called *lexicon* of a meeting recording, i.e. a set $V = \{w_1, \ldots, w_{|V|}\}$, where $|V|$ is number of elements in V containing the list of unique words that can be found in the speech transcription, the slide transcription, or both. The lexicon is used to represent the transcriptions of utterances and slides through vectors $u = (u_1, \ldots, u_{|V|})$, also called *term-histogram vectors*, where u_i is the number of times word w_i appears in the utterance or slide transcription. Such an approach is called *bag of words* and does not take into account the order of the words, but only their frequence.

Given an utterance $u^{(i)}$ and its ASR N-Best list $T = \{t_1, \ldots, t_N\}$, where $t_i = (t_i^1, \ldots, t_i^{|V|})$ is the term-histogram vector representation corresponding to the i^{th} ranking transcription hypothesis, we

	IS1009B	IS1009C	IS1009D	TOTAL
#Slides	16	34	9	59
W_{speech}	6297	4565	4849	15711
W'_{speech}	**2536**	**2046**	**2059**	**6641**
W_{slides}	532	805	230	1567
W'_{slides}	**341**	**541**	**175**	**1057**

Table 1: Word counts in the speech and in the slides, before stopping (W_{speech},W_{slides}) and after stopping (W'_{speech},W'_{slides}).

select as a transcription of $u^{(i)}$, the $t_n \in T$ such that

$$t_n = \arg \max_{0 \leq i \leq N} \sum_{j=-R}^{R} t_i s^{(c+j)} \tag{1}$$

where $s^{(c)}$ is the slide displayed when $u^{(i)}$ is uttered.

2.3 Data Description

Three meetings (IS1009B, IS1009C, IS1009D) were used in the experiments. Of course, they were not part of the training set of the speech recogniser. Each meeting involves 4 participants (mostly non-native english speekers) and has a duration of about 30 minutes. Table 1 gives the number of slides and words (before and after stopping) for each meeting.

3 Results

Recognition results are presented in Table 2. N-Best lists of size $N > 10$ could not be obtained for meeting IS1009D due to the presence of some very short utterances and to lattice pruning during the recognition process. No statistically significant WER variations are observed, and the slight WER changes shown in the table are statistical fluctuations. The following sections provide possible explanations based on both linguistic and statistical arguments.

		Word Error Rate							
		IS1009B		IS1009C		IS10099D		OVERALL	
N	R	GT	OCR	GT	OCR	GT	OCR	GT	OCR
1 BL	0	32.4%		34.7%		32.7%		33.2%	
10	0	32.5%	32.5%	34.1%	34.1%	32.6%	32.7%	33.0%	33.0%
	2	32.5%	32.5%	34.0%	34.1%	32.6%	32.7%	33.0%	33.0%
50	0	32.6%	32.5%	34.1%	34.3%			33.3%	33.3%
	2	32.6%	32.5%	34.2%	34.4%			33.3%	33.3%
100	0	32.7%	32.5%	34.2%	34.4%			33.3%	33.3%
	2	32.7%	32.6%	34.3%	34.5%			33.4%	33.4%

Table 2: Selected rescoring results for different N-Best list sizes and context radii R. BL stands for the baseline performance, GT for the groundtruth transcriptions and OCR for results obtained with the text extracted from the slides through OCR.

3.1 The Zipf's Law

The hypothesis exploited in the present paper was to assume that what is being said during a meeting is correlated with the corresponding slide text. If this assumption is true, one can believe that it is possible to find an approach using the words from the slides to improve the ASR transcription. However, due to the sparse nature of the language, this will generally not be the case. Zipf's law [7] states that a few words occur very often while many others occur rarely. If we rank all the words with respect to the number of times they appear in a corpus, the number n_r of words appearing at rank r will usually be inversely proportional to r:

$$n_r = \alpha / r \tag{2}$$

where α is a constant. In the Wall Street Journal corpus (Figure 3), a frequently used IR benchmark, the 260 most frequent words, corresponding to only 0.3% of the lexicon, represent 50.7% of the word mass (i.e. the total number of words in the corpus). On the other hand, more than 33% of the lexicon terms appear only once in the corpus and correspond to just 0.44% of the word mass. Zipf's law is general and it applies to any text corpus including the case under consideration in this work. More particularly, the word mass from the speech is dominated by a few frequent words. Table 3 shows that the stopwords represent 60% of the word mass, thus at least 60% of the words will not be affected by the rescoring process. Moreover, the only words that could contribute to the improvement of the WER are the terms that are subject specific and co-occur in both speech and slides. Since they are few, their contribution to the WER reduction is limited and does not have a statistically significant impact. On the other hand, the rescoring process might improve the performance of more

	IS1009B	IS1009C	IS1009D	TOTAL
W'_{speech}	40.3%	44.8%	42.5%	42.3%
$\mathrm{d}W'_{speech}$	79.7%	80.1%	79.5%	79.8%
W'_{slides}	64.1%	67.2%	76.1%	67.5%
$\mathrm{d}W'_{slides}$	78.4%	92.1%	77.0%	79.9%

Table 3: Word mass and lexicon (prefix d) percentages after stopping (W' stands for percentages after stopping).

targeted ASR applications, such as *Named Entity* or *Keyword extraction*, that depend only on such a few subject-specific words and that can thus be significantly improved even by correctly classifying few more words than in a baseline system (i.e. a system without rescoring process).

3.2 Pearson's chi square test statistic

From a statistical perspective, we tested whether the distribution of words present in the speech signal is affected by the slides being displayed, i.e. if certain words tend to be said more or less often depending on the slide being displayed. If the word distribution is affected by the slides, then the slides bring additional knowledge that can be exploited in the recognition. We used the Pearson's chi-square test [10] to test the hypothesis of independence between the words being spoken and the slide j displayed. In case of independence we should have:

$$\mathbf{H_0}: \quad p(w_i, s_j) = p(w_i)p(s_j) \tag{3}$$

where $p(w_i, s_j)$ is the probabality of word w_i being said while slide s_j is displayed, $p(w_i)$ is the probability of word w_i being said and $p(s_j)$ is the probability of slide s_j being displayed. The timeline of a meeting was segmented in slots corresponding to the intervals where a single slide j is being displayed, i.e. for J slides, a meeting is segmented into J intervals with boundaries corresponding to the slide transitions. For a vocabulary (as described in Section 2.2) of I words, a table of size $I \times J$

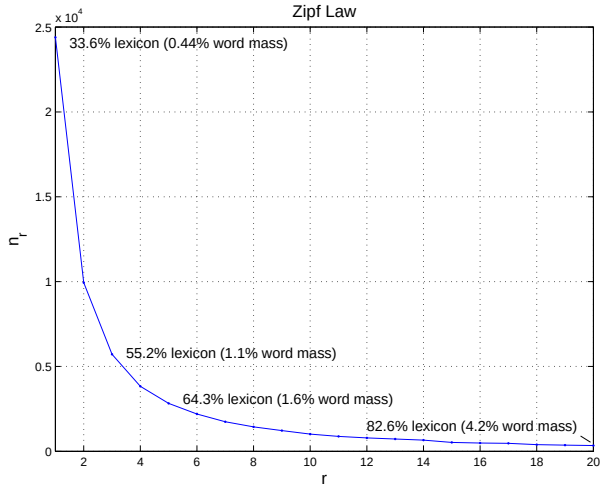

Figure 3: Zipf's law

is populated so that the content a_{ij} of cell (i, j) corresponds to the number of times word i is said during the time interval when the slide j is displayed. Given such table, the Maximum Likelihood (ML) estimation of the above probabilities can be obtained as follows:

$$p(w_i, s_j) = \frac{a_{ij}}{\sum_{ij} a_{ij}}; \quad p(w_i) = \frac{\sum_j a_{ij}}{\sum_{ij} a_{ij}}; \quad p(s_j) = \frac{\sum_i a_{ij}}{\sum_{ij} a_{ij}} \tag{4}$$

Pearson's chi-square test can then be used to verify the statistical independence between the rows and the columns of the table, i.e. the hypothesis $\mathbf{H_0}$. Since in the case under consideration a large number of cells is empty because some words do not occur during certain slides (zero-entries in the table), we applied a smoothing technique. We used the test for all available meetings, for both automatic and manual (groundtruth) transcriptions. In all cases, the hypothesis $\mathbf{H_0}$ of independance between the rows and the columns of the table is confirmed with a confidence of 99%. This result confirms that, even if certain (rare but maybe important) words co-occur with certain slides, the word mass as a whole masks this phenomenon since its distribution is independent of the slides being displayed.

4 Conclusions and Future Work

In this paper, we presented preliminary results on the use of slides content as a means to improve the performance of a LVCSR system applied to meeting recordings. Results obtained by using a N-Best list rescoring technique based on the text extracted from the slides do not show statistically significant WER variations. Only a few words are affected by the rescoring process, which is based on words co-occuring in both streams. These co-occurences are masked by the word mass from the speech, as confirmed by language properties: this is the main reason of the lack of significant improvements. Statistics of the word distribution further validate this hypothesis, as they show no correlation between the words being spoken and the slides. Finally, the provided explanations are independent of the technique used to integrate the information from the slides and thus seem to exclude a problem with the approach adopted in this work.

On the other hand, there are applications like Named Entity or keyword extraction that depend critically on the recognition of the few words that are subject dependent and specific of a certain meeting. Since named entities are often reported on the slides (e.g. author names or acronym definitions) and words co-occuring in slides and speech are potentially good keywords, the approach we proposed in this paper can maybe be more beneficial in such kind of applications.

References

[1] D. Pallett, "A Look at NIST's Benchmark ASR Tests: Past, Present, and Future," in *ASRU*, 2003.

[2] T. Hain and al., "The 2005 AMI System for the Transcription of Speech in Meetings," in *NIST MLMI Meeting Recognition Workshop*, Edinbourgh, GB, 2005.

[3] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 61–69, Sept. 2005.

[4] D. Moore, "The IDIAP Smart Meeting Room," IDIAP-COM 07, IDIAP, 2002.

[5] D. Chen, J.-M. Odobez, and H. Bourlard, "Text Detection and Recognition in Images and Videos," *Pattern Recognition*, vol. 37, no. 3, pp. 595–609, Mar. 2004.

[6] A. Vinciarelli and J-M. Odobez, "Application of information retrieval technologies to presenta tion slides," *IEEE Transactions on Multimedia, to appear*, 2005, IDIAP-RR 05-36.

[7] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[8] F. Walls, H. Jin, S. Sista, and R. Schwartz, "Topic detection in broadcast news," in *Proceedings of the DARPA Broadcast News Workshop*, 1999, pp. 193–198.

[9] E.R. Jessup M.W. Berry, Z. Drmac, "Matrices, vector spaces, and information retrieval," *SIAM Review*, vol. 41, no. 2, pp. 335–362.

[10] R. Christensen, *Log-Linear Models and Logistic Regression*, Springer, 1997.